

Vers l'application de l'apprentissage par renforcement inverse aux réseaux naturels d'attention

Somon, B.; Fermo, A.; Chanel, C.P.C.; Dehais, F.

ANITI / ISAE-SUPAERO

bertille.somon@isae-supraero.fr



Problématique

Définitions Apprentissage par renforcement (inverse), EEG

Travaux antérieurs Oscillations et connectivité "dynamique"

Protocole expérimental Stimuli, méthodes, mesures

Résultats préliminaires

Limites et perspectives

Quelles sont les politiques mises en place par le cerveau humain pour répartir de manière optimale les ressources attentionnelles limitées dont il dispose?

Reinforcement learning (RL)

Les méthodes de RL abordent la question de la maximisation des récompenses futures en associant des états dans l'environnement à des actions. (trad. Hassabis et al., 2017)

- Inspirées de modèles biologiques: *Temporal-Difference (TD) learning*
- Moyen de résolution des **Processus Décisionnels de Markov** (PDM): n-uplet $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma\}$, où:

\mathcal{S} : ensemble d'états;

\mathcal{A} : ensemble d'actions;

$\mathcal{T}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$: fonction de transition d'état telle que $\mathcal{T}(s', a, s) = p(s'|s, a)$;

$\mathcal{R}: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$: fonction de récompense

$\gamma \rightarrow [0, 1[$ est le facteur d'oubli

Définitions

Apprentissage par renforcement (inverse)

- Résolution d'un PDM: cherche une politique (optimale π^*) qui maximise la fonction de valeurs
- RL: Agent interagit avec l'environnement (exploration et/ou exploitation des séquences d'action)
- Évaluation des récompenses obtenues en moyenne et valeur d'une action approchée (modèles estimés ou non)

Définitions

Apprentissage par renforcement (inverse)

- Résolution d'un PDM: cherche une politique (optimale π^*) qui maximise la fonction de valeurs
- RL: Agent interagit avec l'environnement (exploration et/ou exploitation des séquences d'action)
- Évaluation des récompenses obtenues en moyenne et valeur d'une action approchée (modèles estimés ou non)

Mais... (Ng and Russell, 2000)

Peu applicable au comportement humain:

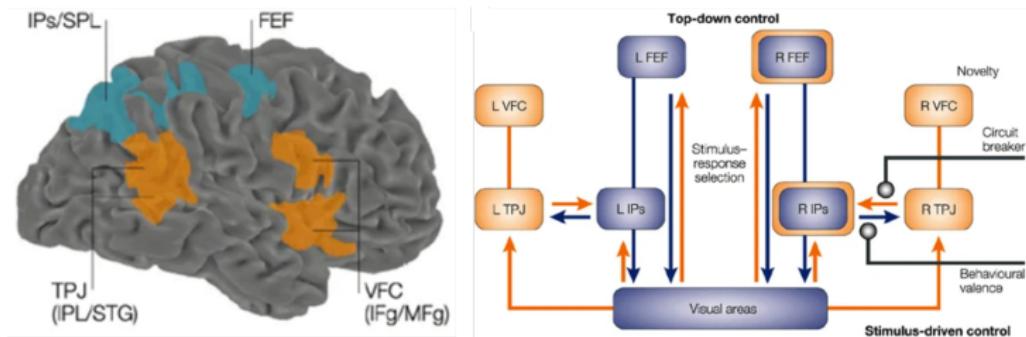
- Supposent des \mathcal{R} fixes et connues
- \mathcal{R} plus complexes que celle généralement admises en RL
- Varient d'un agent à un autre
- Préférable d'inférer \mathcal{R} plutôt que la politique: robustesse, rapidité, transferabilité

Définitions

Attention: exploration vs. exploitation

Deux circuits attentionnels (Corbetta and Shulman, 2002):

- Attention descendante (*top-down*): associée au but; dépend des processus cognitifs en jeu;
 - Attention ascendante (*bottom-up*): associée aux stimulations extérieures; court-circuite le système descendant
- ⇒ Deux réseaux d'activation bien distincts
- ⇒ Varient en fonction de la modalité sensorielle

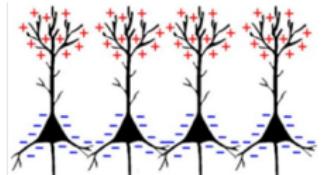


Définitions

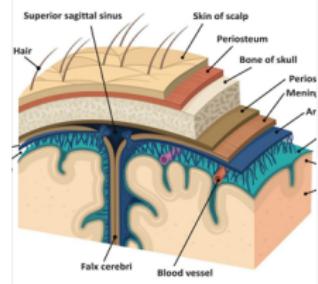
Electroencéphalographie (EEG)

- Mesure de l'activité électrique des couches superficielles du cortex cérébral
 - Synchronisation de populations $\sim 10^4$ neurones/ mm^3
- Macrodipoles tangentiels ou radiaux

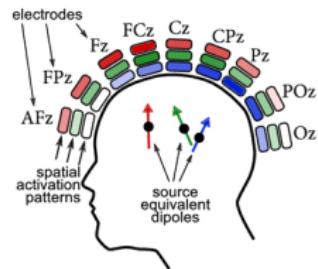
S
sources



W
weighting
coef



X
Recorded
signal



Pour

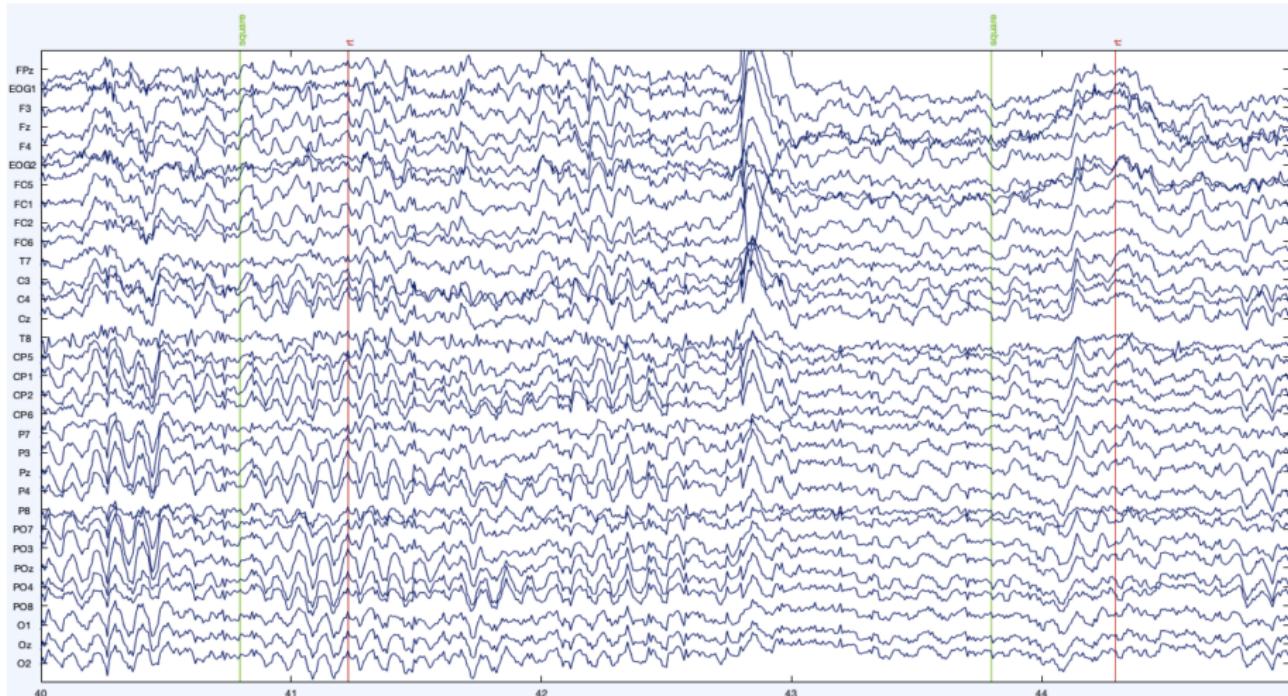
Temporalité cognitive
Mesure directe
Multidimensionnel
Peu encombrant

Contre

Sources superficielles
Mauvais RSB
Incertitude spatiale
Analyses et visualisa-
tion complexes

Définitions

Electroencéphalographie: Activité EEG

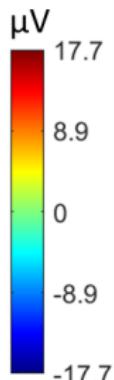
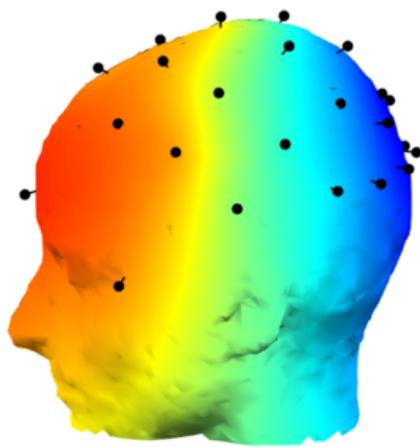


32 électrodes × 5 secondes

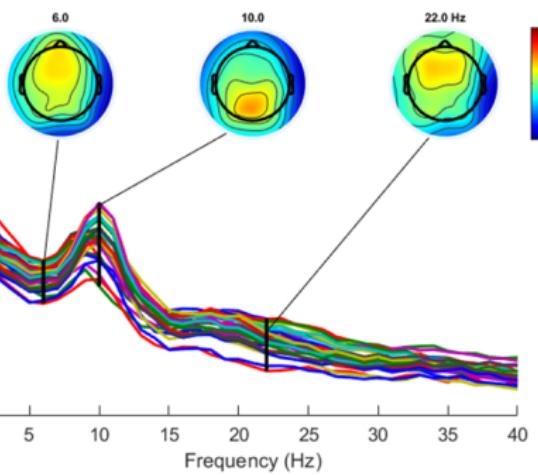
Définitions

Electroencéphalographie: Activité EEG

Potentiels évoqués moyens à 280ms



Log Power Spectral Density $10^* \log_{10} (\mu\text{V}^2/\text{Hz})$

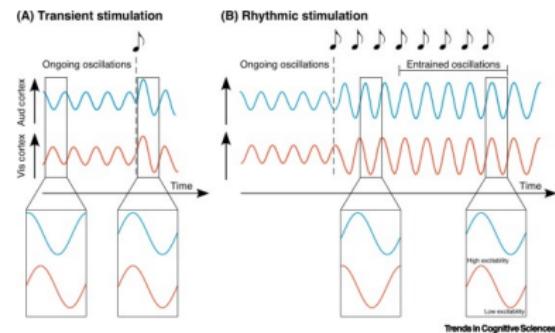


32 électrodes \times fréquences

Problématique

Travaux antérieurs: Bauer et al., 2020

- Regroupement des informations sensorielles se fait uniquement dans les aires multisensorielles ou les aires de haut niveau
- Influences inter-modales aussi très précoce: aires sensorielles primaires
- Positive (amorçage multimodale) ou négative (surdité/cécité inattentionnelle)



- réinitialisation de phase
- entraînement neuronal

- ⇒ Activités dans le domaines **fréquentiel**
- ⇒ Mis en évidence par connectivité dirigée
- Identification des aires actives (*clustering*) et de leurs communications

Problématique

Travaux antérieurs: Bullmore & Sporns, 2009 & Kaminski et al., 2018

Connectivités dirigées: *Directed Transfer Function* (DTF) issue de la causalité de Granger et formulées par modèle AutoRegressif MultiVarié

Modèle MVAR: $X(t) = \sum_{d=1}^p A_{ij}(d)X(t-d) + e(t)$

avec: $X(t)$ l'activité cérébrale aux k électrodes

$A_{ij}(d)$ (taille $k \times k$) les coefficients du modèle d'ordre p

Définit le fonction de transfert du modèle MVAR:

$$\hat{H}_{ij}(f) = \left(\sum_{d=0}^p A_{ij}(d)e^{-2i\pi df\Delta t} \right)^{-1}$$

... et la DTF: influence causale de l'électrode j sur l'électrode i

$$DTF_{j \rightarrow i}(f) = \frac{|H_{ij}(f)|^2}{\sum_{m=1}^k |H_{im}(f)|^2}$$

Problématique

Travaux antérieurs: Bullmore & Sporns, 2009 & Kaminski et al., 2018

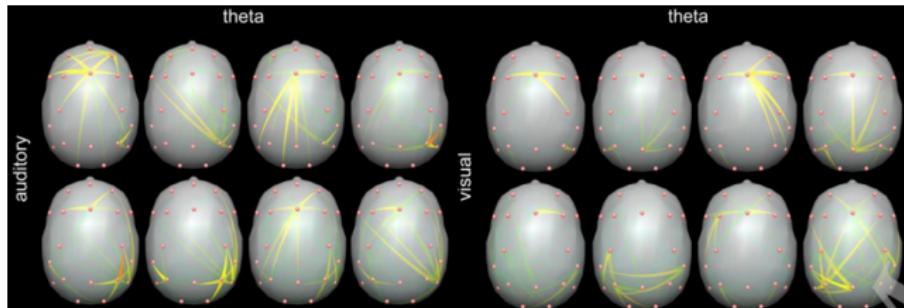
- DTF limite le nombre de connections fallacieuses
- Étendue pour limiter l'impact du dénominateur (dépendant de la fréquence): *full-frequency DTF*
- Mais: affranchissement de l'aspect temporel
 - ⇒ Définition de la *short-time DTF* (sDTF)
 - ⇒ Estime les changements dynamiques dans la propagation de l'activité cérébrale
 - ⇒ Utilisation de courtes fenêtres temporelles

$$\tilde{R}_{ij}(s) = \frac{1}{N_T} \sum_{r=1}^{N_T} \frac{1}{N_S} \sum_{t=1}^{N_S} X_i^{(r)}(t) X_j^{(r)}(t + s)$$

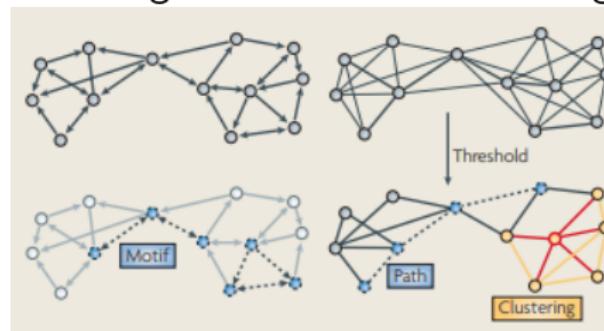
Problématique

Travaux antérieurs: Bullmore & Sporns, 2009 & Kaminski et al., 2018

- Calcul des matrices d'adjacences sur sDTF:



- Seuillage basé sur la théorie des graphes:



Noeuds: électrodes

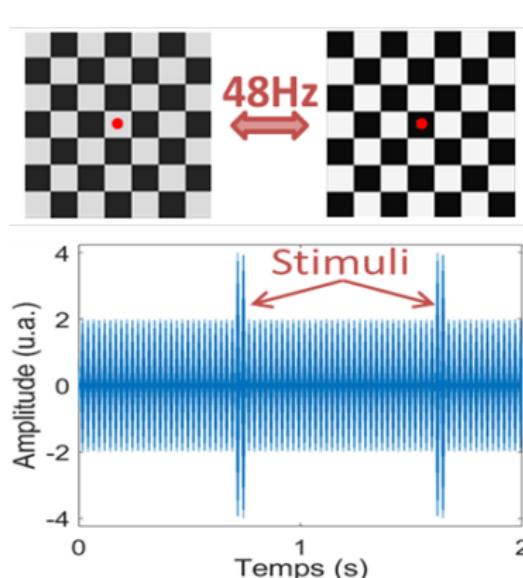
Liens: DTF/sDTF

Métriques: degré des noeuds, densité de connection, longueur des chemins, etc.

Protocole expérimental

Visuo-Auditory Steady State Response Task

Changement d'attention inter-modal



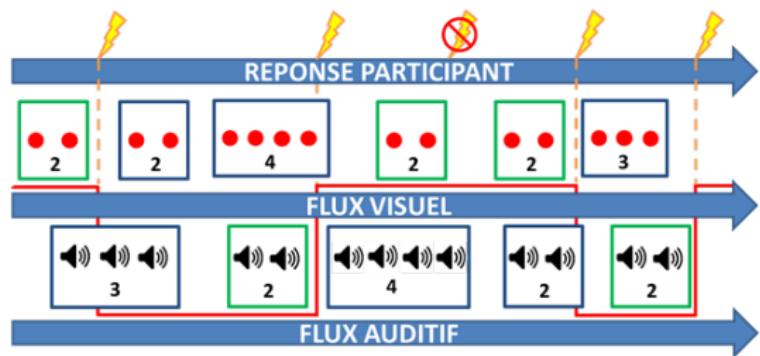
Conditions

Visuelle

Auditive

Visuo-Auditive

Audio-Visuelle



14 participants (7 femmes)

Protocole expérimental

Mesures

- Comportementales: Temps de réactions (s), Taux de bonnes réponses/d'erreurs (%); Taux de *switch*
- Performances: Taux de bonnes réponses, Indice de dominance visuelle
→ Définition experts vs. novices
- Enregistrement EEG continu:

Statiques

Puissance fréquentielle (RESS)
Rapport Signal-Bruit (RESS)
DTF

Dynamiques

Magnitude de la transformée de Hilbert au cours du temps
sDTF

⇒ Définition du vecteur de *features* θ pour chaque état

Connectivité dirigée = Transition $V \rightarrow A$ et $A \rightarrow V$

Protocole expérimental

Modèle

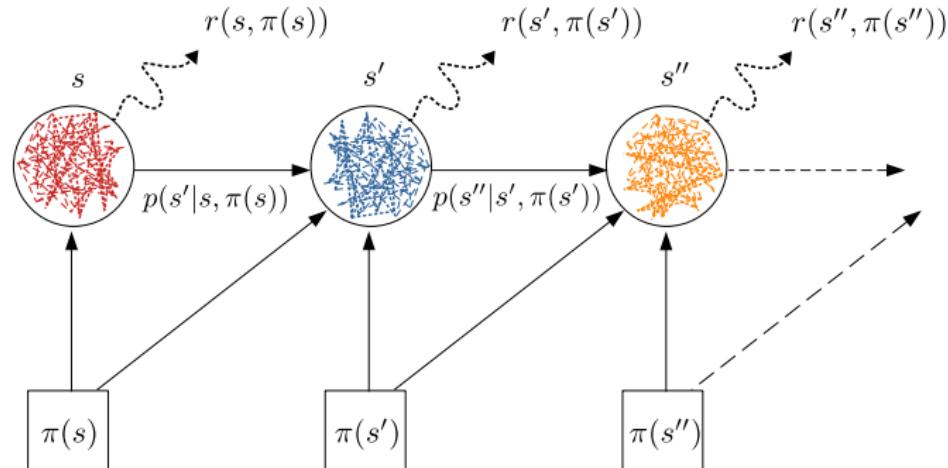
s : état défini par les graphes de connectivité

$\pi(s)$: action du participant

$p(s'|s, \pi(s))$: probabilité de transition

r : récompense engendrée

$$\pi(s) = \arg \max_{a \in \mathcal{A}} Q^\pi(s, a)$$



- ⇒ Modèle MDP dans lequel les états sont ceux définis par l'"état cognitif" du participant à un instant t
- ⇒ Comparaison de participants **experts** (i.e. politique optimale) et de participants **novices**

Protocole expérimental

Modèle

$\mathcal{M} \setminus \mathcal{R}$ où:

\mathcal{S} : ensemble d'états \rightarrow sDTF

\mathcal{A} : ensemble d'actions \rightarrow réponses du participant

$R^*(s, \pi(s)) \in \mathcal{H}_\phi(s, \pi(s))$: espace d'hypothèses de la fonction de récompense à déterminer

On suppose (généralement) que \mathcal{R} est un combinaison linéaire de n features: $r(s, \pi(s)) = \theta^T \phi(s, \pi(s))$

Puis on cherche un θ optimal: θ^* qui explique au mieux la politique réalisée...

... ou (généralement) **minimise par itérations successives** la fonction de perte:

$$\theta^* = \arg \min_{\theta} \mathcal{L}(\mu^{\pi^*}, \mu^{\pi_\theta}) \quad (1)$$

Nous supposons que:

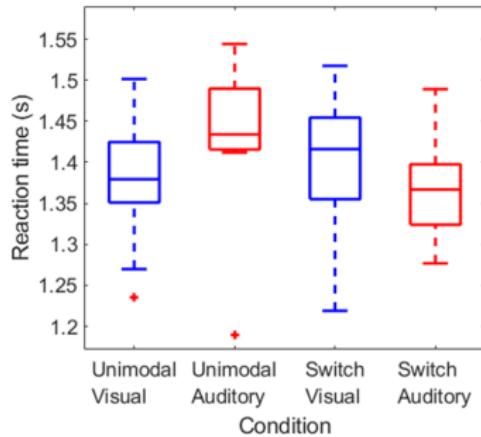
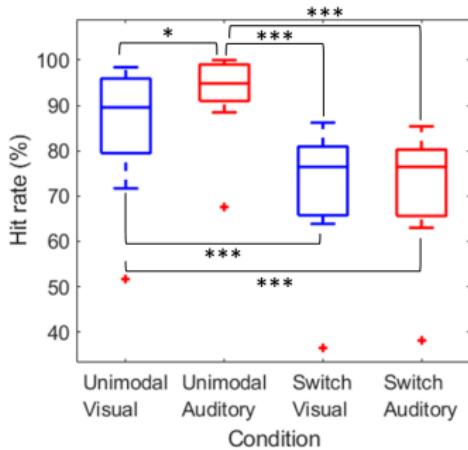
- ① espaces d'états et d'actions discret
- ② $\text{features } \theta$ définis par mesures observées
- ③ Variabilité des fonctions de récompense: il est plus pertinent d'inférer une fonction de récompense stochastique (distribution de probabilité)
- ④ l'expérimentation fourni en entrée des trajectoires de plusieurs (~30aine) d'agents
- ⑤ nécessité d'agréger les données multidimensionnelles

⇒ Algorithme de Babes-Vroman (2011)

Modèle probabiliste qui permet par clustering de définir plusieurs groupes d'agents (e.g. novices et experts) selon la distribution de probabilité sur les fonctions de récompense qui leur est associée

Résultats préliminaires

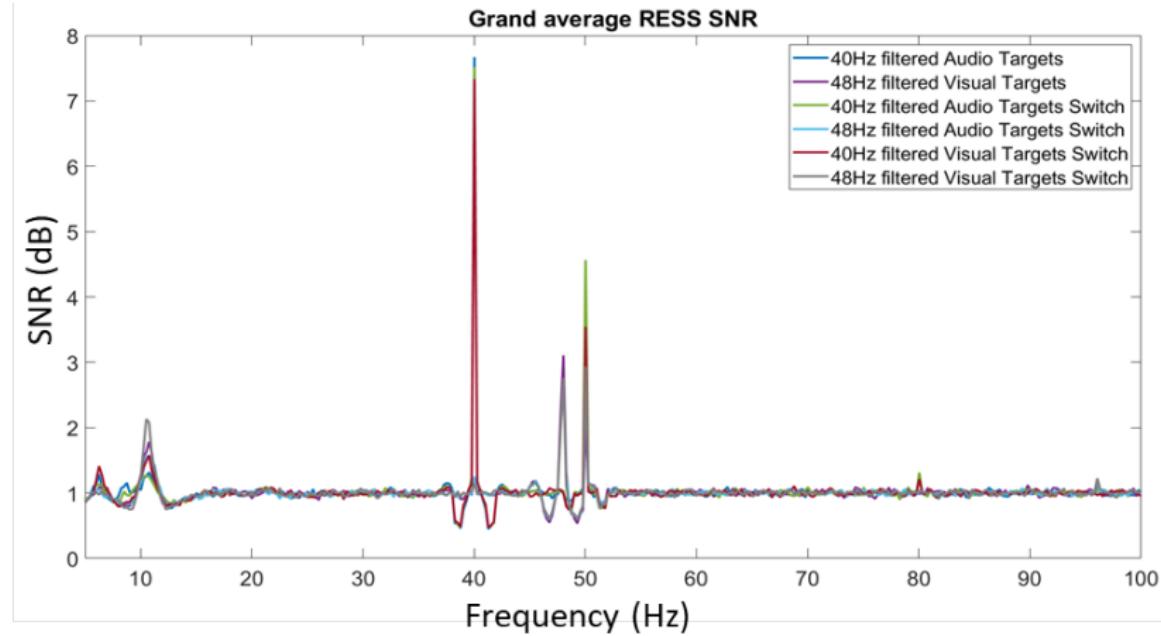
Comportementaux



Analyse de variance à mesures répétées:
condition × modalité

Résultats préliminaires

EEG



Bande alpha, pic à 40Hz (+ harmonique à 80Hz), pic à 48Hz (+ harmonique à 96Hz)

- Tâche très (trop?) simple: pas de prise de décision à proprement parler, tâche de réaction \Rightarrow Modèle MDP?
- Peu d'erreurs en unimodal: pas de comparaison
- Fréquence des steady-state élevée:
 - Meilleure implémentation en condition réelle
 - Incluses dans la bande gamma: difficultés pour le CFC
- Perspectives:
 - Modification de la tâche expérimentale: plus de choix/réponses, plusieurs stratégies
 - Un agent peut suivre plusieurs fonctions de récompense qui évoluent au cours du temps: intégration de cycles dans un modèle bayésien
 - Interfaces Cerveau-Machines
 - Définition d'heuristiques pour les algorithmes d'IA: Attention
 - Formation des populations novices sur la base des états experts